Bayesian meta-analysis for identifying periodically expressed genes in fission yeast cell cycle

Submitted to the Annals of Applied Statistics

BAYESIAN META-ANALYSIS FOR IDENTIFYING PERIODICALLY EXPRESSED GENES IN FISSION YEAST CELL CYCLE

By Xiaodan Fan, Saumyadipta Pyne and Jun S. Liu*

Harvard University, Chinese University of Hong Kong and Broad Institute

The effort to identify genes with periodic expression during the cell cycle from genome-wide microarray time series data has been ongoing for a decade. However, the lack of rigorous modeling of periodic expression as well as the lack of a comprehensive model for integrating information across genes and experiments has impaired the effort for the accurate identification of periodically expressed genes. To address the problem, we introduce a Bayesian model to integrate multiple independent microarray data sets from three recent genome-wide cell cycle studies on fission yeast. A hierarchical model was used for data integration. In order to facilitate an efficient Monte Carlo sampling from the joint posterior distribution, we develop a novel Metropolis-Hastings group move. A surprising finding from our integrated analysis is that more than 40% of the genes in fission yeast are significantly periodically expressed, greatly enhancing the reported 10-15% of the genes in the current literature. It calls for a reconsideration of the periodically expressed gene detection problem.

1. Introduction. Cell division cycle is the concerted sequence of processes by which a cell duplicates its DNA and divides into two daughter cells. Many genes are expressed periodically at a specific stage during the cell cycle when they peak and trough over a certain time range. They are termed as "cell cycle-regulated genes". Here, in the context of mRNA expression studies, we call these "Periodically Expressed (PE) genes". In contrast, other genes are called "APeriodically Expressed (APE) genes". Identification of PE genes is both of theoretical importance because of the need to understand the different mechanisms underlying these genes' involvements in the cell cycle processes, and of practical importance due to the biological links between cell cycle control and many diseases such as cancer (Sherr, 1996; Whitfield *et al.*, 2002; Bar-Joseph *et al.*, 2008).

With the help of the microarray techniques and various cell phase synchronization methods (synchronizing the progression of cells through the stages

*To whom correspondence should be addressed

Keywords and phrases: cell cycle, periodically expressed gene, microarray time series, meta-analysis, fission yeast, *Schizosaccharomyces pombe*, Markov chain Monte Carlo

of cell cycle), researchers have conducted genome-wide time series expression analyses on synchronized cells for various species ranging from fungi to plant to human (Cho *et al.*, 1998; Spellman *et al.*, 1998; Laub *et al.*, 2000; Ishida *et al.*, 2001; Menges *et al.*, 2002; Whitfield *et al.*, 2002; Rustici *et al.*, 2004; Peng *et al.*, 2005; Oliva *et al.*, 2005; Bar-Joseph *et al.*, 2008). Several strategies for identifying PE genes on these data have been developed, such as the fitting of a sinusoidal function (Spellman *et al.*, 1998), clustering techniques (Eisen *et al.*, 1998; Whitfield *et al.*, 2002), the single-pulse model (Zhao *et al.*, 2001), the partial least squares regression approach (Johansson *et al.*, 2003), the average periodogram (Wichert *et al.*, 2004), the linear combination of cubic B-spline basis (Luan and Li, 2004), the random-periods model (Liu *et al.*, 2004), the least square fitting for the periodic-normal mixture model (Lu *et al.*, 2004), the Fourier score combined with p-value of regulation (de Lichtenberg *et al.*, 2005), the robust spectral estimator combined with g-statistic (Ahdesmaki *et al.*, 2005), and the up-down signature method (Willbrand *et al.*, 2005). Zhou *et al.* (2005) applied a Bayesian approach for single experiment data by fixing the period at pre-estimated value. Most of these methods use a set of known PE genes to estimate the cell cycle period prior to testing the periodicity for other genes.

While the previous efforts have often reported positively about the presence of periodic signal in these gene expression data, doubts were raised as to whether such periodic gene regulation was reproducible (Shedden and Cooper, 2002; Wichert *et al.*, 2004) and, by extension, about the identity and count of PE genes discovered by subsequent analyses. One prevalent reason for skepticism is the reliance of many of the studies on *ad hoc* thresholds to classify genes as PE or otherwise. For example, Cho *et al.* (1998) detected the PE genes by visual inspection; Spellman *et al.* (1998) designed a cutoff value based on prior biological knowledge. Another possible reason is that the commonly assumed white noise background model for time series might be too unrealistic to allow correct inference about the identity and count of PE genes (Futschik and Herzel, 2008). Furthermore, all previous approaches were designed for analyzing single time series per gene, which did not allow for an efficient combination of data from multiple experiments and therefore lacked the power to identify a large fraction of all PE genes. Recently Tsiporkova and Boeva (2008) proposed a procedure to combine multi-experiment data based on a dynamic time warping alignment technique, which is potentially useful for analyzing multiple cell cycle data sets if combined with a periodicity detection algorithm. However, the procedure requires each time point within a time series to be aligned to a time point within the other time series, which is not always appropriate when

the lengths of cell cycle period, the sampled time ranges, and the sampling frequencies are all different between experiments.

Recently three independent studies (Rustici *et al.*, 2004; Peng *et al.*, 2005; Oliva *et al.*, 2005) conducted elutriation and cdc25 block-release synchronization experiments to measure genome-wide expression in fission yeast (*Schizosaccharomyces pombe*) cell cycle. The results from these three studies also showed discrepancies with regard to the identity and count of PE genes. They reported 407, 747 and 750 PE genes, respectively, with only 176 genes being common to all three lists. However the availability of 10 genome-wide experiments produced by these three different labs has made the fission yeast currently the organism with the largest cell cycle transcriptome data, which provides us an opportunity to obtain a better understanding of the cell cycle. Marguerat *et al.* (2006) combined the ten data sets from the three studies by multiplying p-values for gene regulation and periodicity from each experiment. They concluded that no more than about 500 PE genes can be reliably identified from the combined data. While observing that well over 1000 fission yeast genes could be periodically expressed and that each study had detected a different subset of these, they attributed the discrepancy to inconsistent gene naming, the use of different data analysis methods, and the use of arbitrary thresholds.

We investigated the PE gene identification problem by employing a Bayesian approach to provide (1) a more realistic and comprehensive model for the cell cycle time series data, and (2) an efficient and rigorous way to combine data from multiple experiments. A hierarchical model together with MCMC computation is used to integrate different sources of variation and correlation into a single coherent probabilistic framework. We applied this approach to integrate the ten genome-wide time series data sets. A striking finding from our analysis is that more than 2000 genes are significantly periodically expressed. This number greatly enhances the count of possible cell cycle regulated genes in the current literature. Most interestingly, our finding can be visualized clearly from Fig 4, which merely displays the *original* data, but with the genes ordered according to our inferred periodicity strength and peaking phase.

2. Materials and methods. In Section 2.1, we describe the cell cycle gene expression data. In Section 2.2, we outline our parametric model for cell cycle gene expression. The Bayesian computation of the model is described in Section 2.3 and Section 2.4. In Section 2.5, we present our strategies for distinguishing PE genes from APE genes based on the model fitting results.

2.1. *Microarray time series data.* We obtained the normalized gene expression data for ten genome-wide experiments by three cell cycle microarray studies (Rustici *et al.*, 2004; Peng *et al.*, 2005; Oliva *et al.*, 2005) from the websites listed in Table 1. For each experiment, a culture of cells are grown and synchronized. A set of microarrays are used to measure gene expressions at selected time points (possibly with technical replication of the microarray). All values were converted to log-ratios with base 2. To make the log-ratios comparable across arrays, we transformed the values for every array separately to set the median log-ratio of each array to zero. Log-ratios from technical replicates, if present, were averaged. Time series with more than 25 percent missing entries were omitted. We unified gene names across the studies based on GeneDB database entries (Hertz-Fowler *et al.*, 2004). The genes without a consistent nomenclature were excluded.

Let Y_{get} denote the gene expression log-ratio at time T_{et} in experiment e for gene g, where $g = 1, \cdots, G$, $e = 1, \cdots, E$, $t = 1, \cdots, S_e$. Here Y_{get} is the observed data; T_{et}, the time of the measurement; G, the total number of genes studied; E, the total number of independent experiments; and S_e, the total number of time points measured in experiment e. The whole data set can be visualized as a G-by-E matrix of time series, where each row corresponds to one gene and each column corresponds to one experiment. If we pool together all filtered data from the ten data sets, we have that G=4994, E=10, and S_e ranges from 18 to 52. A detailed overview of the data is given in Table 1. For illustration, the data for two genes are shown in Fig 1.

2.2. *Model.* We model each time series as a mean curve with additive independent and identically distributed (i.i.d.) Gaussian noise for measured time points. The mean curve is a function of time consisting of a trend component and a periodic component. For the trend component, we uses a linear function along with a truncated quadratic function to model the block-release effect (artifacts introduced by experimental protocols for synchronization; see Lu *et al.* (2004)) and the general trend shown by the time series. We assume a first order Fourier model for the periodic component. A damping term is added to the periodic component to model the cell cycle de-synchronization effect, which implies that the periodic phenomenon eventually disappears as time increases. To model the whole matrix of time series, we assume that the periodic components for all genes within one experiment share the same period, which is equal to the cell division time (i.e. duration between the birth of a cell up to its division into two daughter cells). We further assume that the relative peak time within the cell cy-

cle for every gene is fixed, which allows all genes to share the same phase shift when the periodic components across experiments are compared. More specifically, we assume the following model (M_1) for each time series:

$$Y_{get} = a_{ge}+b_{ge}T_{et}+c_{ge}(\min(T_{et}-d_{ge},0))^2+A_{ge}\cos(\mu_e T_{et}+\psi_e+\phi_g)e^{-\lambda_e T_{et}}+\varepsilon_{get}$$

where

$a_{ge}+b_{ge}T_{et}+c_{ge}(\min(T_{et}-d_{ge},0))^2$: trend component
$A_{ge}\cos(\mu_e T_{et}+\psi_e+\phi_g)e^{-\lambda_e T_{et}}$: periodic component
$\varepsilon_{get}\sim N(0,\sigma^2_{ge})$: i.i.d. noise
a_{ge}, b_{ge}: coefficients of the linear trend of a time series
d_{ge}: ending time of block-release effect of a time series
c_{ge}: magnitude of block-release effect of a time series
σ^2_{ge}: noise level of a time series
A_{ge}: amplitude of periodic component of a time series
μ_e: cell cycle angular frequency, equal to 2π divided by the period of cell cycle of an experiment
ψ_e: experiment-specific phase, which models the phase-shift between two experiments
ϕ_g: gene-specific phase, which decides its peaking time
λ_e: magnitude of the de-synchronization effect of an experiment

For each gene, we use different amplitude parameter A_{ge} for different experiments to account for the effects of different experimental platforms and synchronization techniques. If a gene is not periodic, the fitted amplitude A_{ge} should be close to zero. For such time series, the phase parameter ϕ_g is redundant. To capture different noise levels in different experiments, we specify a hierarchical structure for the noise component by assuming that all σ^2_{ge} from the same experiment share the same inverse chi-square distribution with chosen degree of freedom C_{12} (a constant specified in Appendix A) and unknown hyper-parameters ζ_e:

$$\sigma^2_{ge}|\zeta_e \sim Inv-\chi^2(C_{12},\zeta_e).$$

For convenience, we introduce the following notation:

$$
\begin{aligned}
Y &\equiv \{Y_{get},\ \text{for}\ \ g = 1, \cdots, G; e = 1, \cdots, E; t = 1, \cdots, S_e\}\text{: expression values} \\
\Theta_e &\equiv \{\mu_e, \psi_e, \lambda_e, \zeta_e\}\text{: experiment-specific parameters} \\
\Theta &\equiv \{\Theta_1, \cdots, \Theta_E\} \\
\Phi &\equiv \{\phi_1, \cdots, \phi_G\}\text{: gene phases} \\
\Gamma_{ge} &\equiv \{a_{ge}, b_{ge}, c_{ge}, d_{ge}, A_{ge}, \sigma^2_{ge}\}\text{: time-series-specific parameters} \\
\Gamma_g &\equiv \{\Gamma_{g1}, \cdots, \Gamma_{gE}\} \\
\Gamma &\equiv \{\Gamma_1, \cdots, \Gamma_G\}
\end{aligned}
$$

All variables may be visualized within a gene-by-experiment (i.e., $G \times E$) matrix (Fig 2), which shows their dependence structure. Each row corresponds to a gene-specific parameter ϕ_g and each column represents the set of experiment-specific parameters $(\mu_e, \psi_e, \lambda_e, \zeta_e)$. Each cell of the matrix corresponds to the variables specific to a time series. The gene-specific parameter ϕ_g is the key to integrate the time series for gene g from multiple experiments. Experiment-specific parameters Θ_e are used to pool information across all genes within a particular experiment.

For model comparison, we also introduce the following model (M_0) for APE genes:

$$Y_{get} = a_{ge} + b_{ge}T_{et} + c_{ge}(\min(T_{et} - d_{ge}, 0))^2 + \varepsilon_{get}$$

The only difference between M_0 (null model) and M_1 (alternative model) is the periodic component $A_{ge}\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}}$.

2.3. *Identifiability.* In the M_1 model, the phase parameters ψ_e and ϕ_g are not identifiable because the joint posterior distribution remains the same if we add a constant z to all ψ_e and subtract z from all ϕ_g. This non-identifiability problem can be solved by fixing one of the phase parameters, but the loss of one degree of freedom makes the MCMC algorithm very "sticky" (slow-mixing). Since we only care about the relative values of ψ_e's and ϕ_g's, we solve the problem by assigning a reasonably tight prior distribution to one of the phase parameters and flatter priors to others, and using a transformation group move to improve mixing of the MCMC chain (see Appendix A.3).

For periodic signal fitting, the angular frequency parameter μ_e is usually non-identifiable because a time series with angular frequency μ_e is also a time series with angular frequency μ_e/n for any positive integer n. We avoid this problem by specifying the periodic signal as a damping single sinusoidal

curve and limiting the domain of μ_e to a bounded range. The bound of μ_e is instituted via its prior which is based on our prior knowledge of the cell cycle duration in fission yeast.

2.4. *Bayesian computation.* We estimate all unknown parameters through MCMC simulation of their joint posterior distribution. More specifically, we use a Metropolis-within-Gibbs algorithm to iteratively sample one set of parameters given all the others:

- Step 1: sample experiment-specific parameters Θ_e conditional on Φ, Γ and Y
- Step 2: sample gene-specific parameters ϕ_g conditional on Θ, Γ and Y
- Step 3: sample time series-specific parameters Γ_{ge} conditional on Θ, Φ and Y

The MCMC chain composed of these basic moves suffers from a slow mixing problem caused by strong correlations among some parameters. We can alleviate the problem by parallelizing each of the three steps based on the conditional independence of the parameters. For instance, we can parallelize the sampling of Γ_{ge} from their full conditional distribution since they are independent of each other given Θ, Φ and Y. When some parameters are highly correlated in their joint distribution, single-component moves cause very slow-mixing. To cope with this problem, we designed a new sampler called Metropolized independence group sampler (MIPS) by combining the ideas of grouping (Liu *et al.*, 1994) and Metropolized independence sampler (Hastings, 1970; Liu, 1996, 2001). The key idea is to update the whole subset of correlated variables simultaneously independent of the current state using a sequential proposing procedure. MIPS moves are inserted to the main Metropolis-within-Gibbs iteration. The details of the MCMC implementation are given in Appendix A.

2.5. *Strategies for discerning PE genes from APE genes.* We used three statistics to judge which genes are PE ones. Among them, Bayesian Information Criterion is used to compare the fitting of model M_1 with that of model M_0, both to real data. The other two statistics measure the periodicity by comparing the fitting of M_1 model to the real data with that to the permuted data or the data simulated from the M_0 model.

2.5.1. *Permutation test.* Since we fit model M_1 to every gene, even the APE genes are modeled with experiment-specific parameters Θ that are primarily determined by PE components. Therefore, to examine the effect of our Bayesian model fitting procedure on APE genes, we generate background

data by permuting each time series for every gene in the real data, which destroys any periodic pattern therein. We run the same MCMC algorithm to fit the M_1 model to the background data set by fixing all experiment-specific parameters Θ at the posterior mode obtained from the MCMC run for the real data.

2.5.2. *Simulation from the null model.* One problem of using the permutation data as background control is that the permuted time series do not capture the intrinsic autocorrelation of the measured time series, which exists even if it is not periodically expressed. For example, many time series in the real data show a general trend without oscillation, which may be a result of the gene's response to the perturbation caused by synchronization techniques. To accommodate this possible bias, we generate a second data set from the M_0 model. Compared to the permuted time series, M_0 explains the autocorrelation in the time series by a mean curve. We run the same MCMC algorithm to fit M_0 to all genes in the real data.

We simulated from the M_0 model a data set of similar size and structure as the combined real data set. All parameters are simulated from their corresponding prior distributions. Both M_1 and M_0 are fitted to this simulated data set. While fitting M_1, we fix all experiment-specific parameters Θ at the posterior mode obtained from the MCMC run for the real data.

2.5.3. *Model comparison.* One approach for discerning PE genes from APE genes is to use permuted data or data simulated from the null model as background control, and to fit the M_1 model to both the real data and the background data. The fitting of the background data is then used to determine a threshold for the desired false positive rate (FPR). Another approach is to fit both models M_1 and M_0 to the real data, and then do the classification based on a comparison of the fitness of the models. Various information criteria can be used for this task, such as Akaike's Information Criterion (AIC) (Akaike, 1973), Bayesian Information Criterion (BIC) (Schwarz, 1978), and Deviance Information Criterion (DIC) (Spiegelhalter *et al.*, 2002), to just name a few.

A full Bayesian alternative to our approach here is to introduce a latent variable I_g for each gene to indicate whether it comes from M_1 or M_0. Then, the reversible-jump strategy (Green, 1995) can be used to build a MCMC sampler to traverse the joint space of the latent indicators and model parameters. But due to the global nature of many parameters in our model, this approach is computationally extremely expensive. Additionally, the results so obtained may be too sensitive to our model assumptions. Thus, we feel that using randomization and null model approaches in the spirit of poste-

rior predictive model checking (Gelman *et al.*, 1996) provides a more robust detection of PE genes.

2.5.4. *Statistics for periodicity.* We use multiple gene-specific statistics to measure the periodicity of a gene. Based on the fitted parameter values for the M_1 model, we define the gene-specific Signal-to-Noise Ratio (SNR) as the relative strength of the fitted periodic component compared to the noise level:

$$SNR_g = \sum_{e=1}^{E} \frac{\sum_{t=1}^{S_e} \{A_{ge} \cos(\mu_e T_{et} + \psi_e + \phi_g) e^{-\lambda_e T_{et}}\}^2}{\sigma_{ge}^2}.$$

The SNR statistic combines periodicity information for a gene from every experiment in terms of the amplitude of its periodic component. For each gene, we calculate SNR for each iteration of the MCMC chain, and then summarize the posterior samples of SNR using the 2.5th percentile, the 97.5th percentile, and the mean. Genes with higher SNR values are more likely to be periodically expressed. We also use the fitted phase to measure periodicity from the fitted parameters of the M_1 model. More specifically, we use the length of the 95% central posterior interval (denoted by LPI) of a gene's relative phase $\phi_g + \psi_1$ (ψ_1 is chosen arbitrarily since only the difference of relative phase matters) as one of the periodicity measures. Genes with a higher LPIs are less likely to be periodic either because their periodic components are too weak or their multiple time series might show inconsistent peaking time within the cell cycle.

We use Bayesian Information Criterion difference (BIC^{01}) to measure periodicity based on the fitted posterior modes of the two models. Let L_g^0 and L_g^1 denote the likelihood values for gene g at the posterior mode of the parameters for models M_0 and M_1, respectively. The model comparison criterion BIC^{01} is defined as $BIC_g^{01} = 2log(L_g^1) - 2log(L_g^0) - (k_1 - k_0)log(N)$, where N is the number of observed data points for the gene, k_0 and k_1 are the number of free parameters in models M_0 and M_1, respectively. A gene with positive BIC^{01} value prefers model M_1 to M_0. Genes with higher BIC^{01} values are more likely to be periodically expressed.

3. Results and discussion.

3.1. *Model fitting check.* The MCMC chain on the entire real cell cycle data converged in approximately 2000 iterations. The autocorrelation function of the posterior probabilities from each chain showed that the MCMC algorithm is efficient in terms of effective sample sizes after burn-in. The details of the model fitting diagnosis are given in the supplemental materials of

this paper. Fig 3 displays the posterior distribution of the cell cycle length $2\pi/\mu_e$ for each of the ten experiments. After convergence, the experiment-specific parameters Θ showed little variation, i.e., their marginal posterior distributions had very small variance compared to their ranges. Based on the posterior mode determined from the MCMC chain, we calculated the residue of each time series. The autocorrelation analysis of the residue showed that by fitting M_1 to the data, the autocorrelation was reduced to the level comparable to those of i.i.d. noise. Comparison of variance reduction between the real and the permuted data suggested that the M_1 model explained a significant amount of variance for most of the genes showing significant autocorrelation in their time series.

3.2. *Number of periodically expressed genes.* We ranked all genes in the order of decreasing posterior mean SNR value. Thus, highly ranked genes are more likely to be periodically expressed. We then stratified this sorted list into 6 groups, re-ordered each group according to the fitted peaking time. Fig 4 shows the whole sorted data set. Strikingly, a periodic pattern stands out for all gene groups after simply reordering them (note that these are simply rearranged original data). The pattern is clear and consistent across all experiments for the top 2000 genes, which suggests that about 40% of all genes in the organism could be periodically expressed. The pattern is still strong for genes in the range 2001-3500. We can even observe periodicity among the remaining genes shown in the bottom group, which however is comparable to the top ranking "genes" in the permuted data.

For a comparison with the result from traditional clustering methods, the microarray clustering software Cluster (Eisen *et al.*, 1998) was used to group genes with similar gene expression. A heatmap similar to Fig 4 is included in the supplemental materials of this paper. Compared to the ubiquitous periodic pattern in Fig 4, only several small clusters with visible periodic pattern may be observed from the hierarchical clustering result.

We used two approaches to test whether the visual periodic pattern in Fig 4 is statistically significant. The first approach compares the fitting of M_1 model to the real and background data, i.e., the permuted data or the data simulated from the M_0 model. Two statistics are used to measure the periodicity for this approach. The SNR statistic measures the amplitude of the periodic component, while the LPI statistic measures the uncertainty of the relative phase of every gene. Fig 5 and Fig 6 show the estimated posterior densities of these measures. The curves from the background data provide a null distribution for the corresponding statistic, from which we can estimate FPR for any given threshold. The clear separation of the posterior

densities for the real and background data suggests that a lot of genes show a periodic pattern that is stronger than i.i.d. noise or M_0 data. For example, by comparing the LPI curves of the real and permuted data in Fig 6, we can claim 3086 PE genes for FPR=0.002, corresponding to about 10 false positives. Similarly, by comparing the posterior mean SNR values of the real and permuted data in Fig 5, we can claim 3599 PE genes for FPR=0.002. The number of claimed PE genes when using the simulated data from the M_0 model as background control is similar. For instance, the comparison of the posterior mean SNR densities yields 3414 PE genes for FPR=0.002, and that of the LPI densities yields 3036 PE genes for FPR=0.002.

The second approach compares the fitting of the two models M_1 and M_0, both using the real data. We used BIC as the model comparison criterion. As shown in Fig 7, almost all BIC^{01} values from the permuted data as well as the simulated data from the M_0 model are smaller than zero. For the real data, we can claim 2003 PE genes from the combined analysis by using zero as the threshold for BIC^{01}. Corresponding to this threshold, the permuted data will only produce one false positive PE gene, corresponding to FPR=0.0002.

The results of these three statistics are summarized in Table 2. Here we used the permuted data as background control. The average Spearman correlation between pairs of the statistics is 0.87, suggesting that the three statistics are highly consistent in ranking the genes' periodicity. The approaches based on permutation control (SNR, LPI) made more significant claims than the model selection approach. Overall, we obtained a list of 1898 significant PE genes that are claimed by all the three statistics.

3.3. *Performance comparison.* To evaluate the performance of identifying PE genes, we defined a benchmark set as the union set of the list of PE genes derived from small-scale experiments (Marguerat *et al.*, 2006) and a core set of genes whose periodic regulation is conserved between budding yeast and fission yeast (Lu *et al.*, 2007). The resulting benchmark set consists of 162 genes. We used this benchmark set to compare our method with the method used by Marguerat *et al.* (2006).

The statistic used for gene classification by Marguerat *et al.* (2006) is a score calculated from a p-value of regulation and a p-value of periodicity. When combining multiple experiments for gene classification, they multiplied the p-values from individual experiments to get a total p-value of regulation and a total p-value of periodicity. To estimate the FPR of their statistic, we calculated the scores for the permuted data. For our method, we use the SNR statistic for gene classification.

Fig 8 shows the performance of the our SNR statistic and Marguerat *et al.*'s score on both the combined data (all experiments) and the Exp1 data (a single experiment) in the form of ROC curves. For any given FPR value, we estimate the threshold of a statistic from the permuted version of the data. The corresponding false negative rate (FNR) is estimated by the fraction of the genes in the benchmark set that are classified as APE gene according to this threshold. When applied on the data from a single experiment (Exp1 data), the SNR statistic apparently outperforms Marguerat *et al.*'s score. The gain of statistical power at the single experiment level could be due to our explicit modeling of the trend component and the de-synchronization effect, which makes our model more realistic for the cell cycle time series. When comparing their performances on the combined data, it seems that the SNR statistic increases the statistical power over Marguerat *et al.*'s score significantly. This is due not only to a more realistic model for single time series, but also to our approach of the Bayesian meta-analysis. Instead of combining the p-values from individual experiments, we model multiple experiments simultaneously so as to borrow information across experiments.

Fig 8 indicates that the same statistic performed better at discerning PE genes with the combined data than with the data from a single experiment. This is also true when comparing the performances of a statistic using the overall combined versus that using any subset of the experiments. The detailed information is given in Table 2 in the supplementary material. This is natural because any subset contains less information than the full combined data; but on the other hand, it also indicates that each experiment captured some information about genes' periodicity during cell cycle.

3.4. *Subset analysis.* To compare three individual studies (Rustici *et al.*, 2004; Peng *et al.*, 2005; Oliva *et al.*, 2005) and different experimental techniques, we used the same method for the combined data set to fit model M_1 to all three individual data sets, also the two collections of experiments using different synchronization techniques (elutriation or cdc25 block-release). We first determined the 95% posterior interval of the SNR statistic for each gene to account for the uncertainty of its SNR estimate. Then for comparison of all the subsets at the same significance level, we claim a gene to be PE if its posterior mean SNR value is above the upper 97.5% posterior limits of the SNR of at least 4984 (out of 4994) permuted "genes". For the combined data, we thus claimed 2032 PE genes. Fig 9A and Fig 9B show the overlap of the results from our subset analyses. Fig 9C shows the overlap of the original results from the three individual studies. There are 976 genes which are reported as PE by our combined analysis but not by any of the three

original studies. Supporting evidences for these genes are included in the supplementary material.

Similar to Fig 9C, the discrepancy about the count and identity of PE genes exists between individual data sets (Fig 9A) and across synchronization techniques (Fig 9B) although we have unified the whole analysis procedure. Therefore instead of attributing the discrepancy between the subsets to inconsistent gene naming or use of different analysis methods or arbitrary thresholds (Marguerat *et al.*, 2006), we suggest that the cause is intrinsic to the data. It also shows that most genes in the discrepant part show significant periodicity in the combined analysis. The combined analysis also captured many genes which can not be detected by subset data analysis. Combined with the benchmark analysis, we observed that 5 out of the 40 benchmark genes whose periodicity have been confirmed by small-scale experiments (Marguerat *et al.*, 2006) were missed by all three original studies as well as our combined analysis. On the other hand, 6 out of the 92 core environmental stress response genes with known function (Chen *et al.*, 2003) were claimed as periodically expressed by all three original studies as well as by our combined analysis, suggesting that their periodic signal is clear to all methods. Possibly, the periodicity measure for widely used positive or negative benchmark sets are not quite accurate.

To investigate the discrepancy between different subsets, we systematically tested these subsets' pairwise reproducibility using the posterior mean SNR values. If it is true that the genes have an intrinsic order in terms of periodicity and all individual data sets are of similar quality in revealing this ordering information, the periodicity measures across pairs of subsets should be consistent. Each data set yields a SNR vector measuring the periodicity of all genes. The key idea is to check whether the Spearman correlation of the two SNR vectors is still significant after removing genes which are top ranked in both vectors. The details are shown in Fig 10. After removing the 847 genes that are highly ranked by both Peng *et al.* and Oliva *et al.*, the remaining genes' SNR values from these two data sets show no positive Spearman correlation at the significance level of 0.05. This sets the number of reproducible genes supported by these two data sets (5 experiments) to 847. This same count increases to 934 for Rustici *et al.* versus Peng *et al.* (7 experiments), and to 1008 for Rustici *et al.* versus Oliva *et al.* (8 experiments). The increasing of reproducible genes is consistent with the increase in the size of data involved in comparison. The number further increases to 1554 when comparing elutriation experiments with cdc25 experiments. This suggests that although the number of reproducible genes is less than the number of PE genes suggested by the combined analysis, the reproducibility

is improved by including more data in the comparison or by partitioning the data according to experiment technique.

To explain the above subset discrepancy, possible flaws in the benchmark sets, and the high number of significant genes in the combined analysis, we hypothesize a network-based dynamics for the cell cycle process. For instance, periodic signals from transcription of key cell cycle-regulated genes propagate through the relevant downstream regulatory networks of the organism potentially targeting a considerable number of genes. Thus, depending on the status of the network, these genes may show an observable periodic pattern under one condition, and be too weak to detect under another condition. As a consequence of the combined effect of the variation in periodicity and experimental noise, each study could capture a different subset of the PE genes. The difference of the cell cycle length shown in Fig 3, which could not be explained solely by microarray platform difference, is a further evidence of such variation. For example, the cell cycle lengths in the posterior mode for the two cdc25 experiments in Rustici *et al.* are 135 and 138 minutes. While in Oliva *et al.* and Peng *et al.*, this number increases to 164 and 173 minutes, respectively. Although they are using the same synchronization technique on the same organism, subtle environmental or physiological differences have changed the speed of the cell cycle oscillation. Therefore, it may have also changed relative amplitudes of oscillation of the genes leading to overall ranking discrepancy.

4. Conclusion. In spite of the rapid rise in the number of microarray experiments, many of which address related issues, a systematic meta-analysis of such data is rarely attempted. We conducted a meta-analysis of ten fission yeast cell cycle genome-wide time-series experiments with a model-based Bayesian approach. Compared to other methods, key features of our model include the fixed relative phase of the peaking time of the genes across all experiments (e.g., a gene will peak 10 degrees earlier than another gene in an experiment if and only if the same happens in another experiment) and a flexible amplitude for periodic components. Our approach does not require training sets to estimate important global parameters such as the period of cell cycle, but to infer them from all the data. Notably, our parametric approach deals with phase shift, signal amplitude difference, noise level difference and de-synchronization automatically. Despite the high dimensionality, the implemented MCMC chain mixes well with the help of global moves. The residual analysis shows that our model fits the data well.

A striking finding of our analysis is that more than 2000 genes are significantly periodically expressed, which accounts for approximately 40% of all

the genes in the fission yeast genome. The subset analysis suggests that this number may increase with more data included. This enhances greatly the current knowledge of only 10-15% of all fission yeast genes that are reported as periodically expressed during the cell cycle. Interestingly, genome-wide oscillation has also been reported by recent studies on other cyclic phenomena in the cell, such as the metabolic cycle and circadian periodicity (Klevecz *et al.*, 2004; Tu *et al.*, 2005; Ptitsyn *et al.*, 2007). Clearly, certain amount of influence of the global cell cycle processes on most genes in the genome, in particular in unicellular organisms such as fission yeast, cannot be ruled out. For instance, the folding and unfolding of chromosomes over the course of cell cycle will have genome-wide incidental effect on transcription. However, earlier studies concede that limited ability to distinguish precisely the weakly periodic oscillations from prevalent microarray noise only allowed conservative estimates of PE genes. By explicitly modeling periodic and non-periodic components, and different sources of variation and noise, our model-based approach helps to overcome this long-standing limitation. The resulting list of more than 2000 PE genes would allow the researchers to cast a much wider and deeper net for cell cycle regulated genes that can lead to investigation of novel or relatively less known gene modules and networks involved in the machinery of cell cycle regulation.

It should be noted that the key idea behind our model is rather general. It can be applied to detect periodic patterns where the amplitude is noisy but the patterns are nonetheless consistent across different experiments. The data can be any collection of time series. A study of cell cycle data from other species such as the budding yeast, mouse, human, etc, using the proposed method can be of immediate interest.

One possible way to improve the current method is to employ a more robust error model, using for example t-distributions instead of Gaussians for the noise term (Hampel *et al.*, 1986; Lange *et al.*, 1989). But as a price to pay, the computational complexity may be increased substantially. It should be noted that, as stated in Section 2.5.3, alternative Bayesian model selection methods may also applied to this problem. For example, Green (1995) provides a way to perform joint model selection and parameter estimation via reversible jump MCMC. It may be applicable to this problem if the efficiency of reversible jump MCMC moves can be improved significantly. The methods proposed by Chib (1995) and Chib and Jeliazkov (2001), which estimate the marginal likelihood of the data under a model, may also be a worthwhile direction to explore.

APPENDIX A: MCMC IMPLEMENTATION

A.1. Prior distribution. We assigned reasonably diffuse but still proper prior distributions for all parameters:

$$\begin{aligned}
a_{ge} &\sim N(0, C_1) \\
b_{ge} &\sim N(0, C_2) \\
c_{ge} &\sim N(0, C_3) \\
d_{ge} &\sim Unif(0, C_4) \\
A_{ge} &\propto Exp(rate = C_5),\ 0 \le A_{ge} < C_6 \\
\mu_e &\sim Unif(C_7, C_8) \\
\psi_1 &\propto N(0, C_9),\ -\pi \le \psi_1 < \pi \\
\psi_e &\propto N(0, C_{10}), e = 2, \cdots, E,\ -\pi \le \psi_e < \pi \\
\phi_g &\sim Unif(-\pi, \pi) \\
\lambda_e &\sim Unif(0, C_{11}) \\
\sigma^2_{ge}|\zeta_e &\sim Inv - \chi^2(C_{12}, \zeta_e) \\
\zeta_e &\sim Exp(C_{13})
\end{aligned}$$

The constants in the prior distributions are assigned correspondingly, making use of our prior knowledge: $C_1 = 1, C_2 = 0.005^2,\ C_3 = 0.0001^2,\ C_4 = 500,\ C_5 = 10,\ C_6 = 10,\ C_7 = 2\pi/180,\ C_8 = 2\pi/120,\ C_9 = 0.2^2,\ C_{10} = 1^2,\ C_{11} = 0.006,\ C_{12} = 4, C_{13} = 50.$

A.2. Posterior distributions and Metropolis-within-Gibbs. We can write down the joint distribution of the data and parameters as:

$$\begin{aligned}
p(Y, \Theta, \Phi, \Gamma) =& p(Y|\Theta, \Phi, \Gamma)p(\Theta, \Phi, \Gamma) \\
=& [\prod_{g=1}^{G}\{\prod_{e=1}^{E}\langle\prod_{t=1}^{S_e} p(Y_{get}|a_{ge}, b_{ge}, c_{ge}, d_{ge}, A_{ge}, \sigma^2_{ge}, \phi_g, \mu_e, \psi_e, \lambda_e)\rangle \\
& p(a_{ge})p(b_{ge})p(c_{ge})p(d_{ge})p(A_{ge})p(\sigma^2_{ge}|\zeta_e)\}p(\phi_g)] \\
& \langle\prod_{e=1}^{E} p(\mu_e)p(\psi_e)p(\lambda_e)p(\zeta_e)\rangle
\end{aligned}$$

We assume that all missing data are missing completely at random, so their corresponding components are simply omitted from this expression.

Again, we introduce the following symbols for convenience:
$D_{get} \equiv Y_{get} - a_{ge} - b_{ge}T_{et} - c_{ge}(\min(T_{et} - d_{ge}, 0))^2$
$R_{get} \equiv D_{get} - A_{ge}\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}}$

$$X_{get} \equiv (1,\ T_{et},\ [\min(T_{et} - d_{ge}, 0)])$$

$$X_{ge} \equiv \begin{pmatrix} X_{ge1} \\ \vdots \\ X_{geS_e} \end{pmatrix}$$

$$Z_{get} \equiv Y_{get} - A_{ge}\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}}$$

$$Z_{ge} \equiv \begin{pmatrix} Z_{ge1} \\ \vdots \\ Z_{geS_e} \end{pmatrix}$$

$$V \equiv \begin{bmatrix} \frac{1}{C_1} & & \\ & \frac{1}{C_2} & \\ & & \frac{1}{C_3} \end{bmatrix}$$

From the joint distribution, we can get all full conditional posterior distributions:

$$\begin{pmatrix} a_{ge} \\ b_{ge} \\ c_{ge} \end{pmatrix} | rest \sim N((\frac{X_{ge}^T X_{ge}}{\sigma_{ge}^2} + V)^{-1}\frac{X_{ge}^T Z_{ge}}{\sigma_{ge}^2},\ (\frac{X_{ge}^T X_{ge}}{\sigma_{ge}^2} + V)^{-1})$$

$$p(d_{ge}|rest) \propto \tfrac{1}{C_4}\exp\{-\tfrac{\sum_{t=1}^{S_e} R_{get}^2}{2\sigma_{ge}^2}\}$$

$$A_{ge}|rest \propto N(\mu, \sigma^2),\ 0 \le A_{ge} < C_6$$

where:

$$\mu = \frac{\sum_{t=1}^{S_e}\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}} D_{get} - \sigma_{ge}^2 C_5}{\sum_{t=1}^{S_e}\{\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}}\}^2}$$

$$\sigma^2 = \frac{\sigma_{ge}^2}{\sum_{t=1}^{S_e}\{\cos(\mu_e T_{et} + \psi_e + \phi_g)e^{-\lambda_e T_{et}}\}^2}$$

$$p(\mu_e|rest) \propto \tfrac{1}{C_8 - C_7}\prod_{g=1}^{G}\prod_{t=1}^{S_e}\exp\{-\tfrac{R_{get}^2}{2\sigma_{ge}^2}\},\ C_7 \le \mu_e < C_8$$

$$p(\psi_e|rest) \propto C_9^{-0.5}\prod_{g=1}^{G}\prod_{t=1}^{S_e}\exp\{-\tfrac{R_{get}^2}{2\sigma_{ge}^2} - \tfrac{\psi_e^2}{2C_9}\},\ -\pi \le \psi_e < \pi,\ \text{for } e = 1$$

$$p(\psi_e|rest) \propto C_{10}^{-0.5}\prod_{g=1}^{G}\prod_{t=1}^{S_e}\exp\{-\tfrac{R_{get}^2}{2\sigma_{ge}^2} - \tfrac{\psi_e^2}{2C_{10}}\},\ -\pi \le \psi_e < \pi,\ \text{for } e = 2, \cdots, E$$

$$p(\phi_g|rest) \propto \prod_{g=1}^{G}\prod_{t=1}^{S_e}\exp\{-\tfrac{R_{get}^2}{2\sigma_{ge}^2}\},\ -\pi \le \phi_g < \pi$$

$$p(\lambda_e|rest) \propto \prod_{g=1}^{G}\prod_{t=1}^{S_e}\exp\{-\tfrac{R_{get}^2}{2\sigma_{ge}^2}\},\ 0 \le \lambda_e < C_{11}$$

$$\sigma_{ge}^2 \sim Inv - \chi^2(S_e + C_{12},\ \tfrac{C_{12}\zeta_e + \sum_{t=1}^{S_e} R_{get}^2}{S_e + C_{12}})$$

$$\zeta_e \sim Gamma(\tfrac{C_{12}}{2}G + 1,\ \tfrac{C_{12}}{2}\sum_{g=1}^{G}\tfrac{1}{\sigma_{ge}^2} + C_{13})$$

For conditional distributions which we only know up to a normalization constant, we used the Metropolis-Hastings algorithm to draw samples. When fitting M_0 model to a gene, the full conditional distribution of its parameters

can be obtained by simply replacing all A_{ge} with zero in the corresponding full conditional distribution from M_1.

A.3. Advanced MCMC moves for better mixing. Besides the basic Metropolis-within-Gibbs iteration, we insert the following moves to perturb the MCMC chain in order to help it traverse faster through the high dimensional space where there are many local modes and strong correlations among a group of parameters.

- Phase parameters ψ_e and ϕ_g are not identifiable in model M_1 because the joint posterior distribution is invariant if we add a value to all ψ_e and subtract the same value from all ϕ_g. One way to solve this non-identifiability problem is to fix one of them, but it appears that the loss of one degree of freedom makes the chain very sticky, i.e., slow to converge. As an alternative, we assign zero-centered normal prior distributions to all ψ_e, and use a transformation group move (Liu and Wu, 1999; Liu and Sabatti, 2000; Liu, 2001) to improve mixing of the MCMC sampler. Specifically, we first propose a move by adding a random number z to all ψ_e and subtracting z from all ϕ_g, and then use the Metropolis-Hastings rule to accept or reject this move. Since we only care about the relative phases of genes and experiments, we use $\phi_g + \psi_1$ as gene's relative phase and $\psi_e - \psi_1$ as the phase for an experiment.
- When a gene violates the assumption that its peaking time in the cell cycle relative to all other genes is fixed across different experiments, its multiple time series will show inconsistent phases, which leads to multiple modes for its phase parameter ϕ_g and amplitude parameters A_{ge}. It is difficult to get out of this kind of local mode by updating ϕ_g and A_{ge} separately and locally. We combine the idea of grouping (Liu *et al.*, 1994) and Metropolized independence sampling (Hastings, 1970; Liu, 1996, 2001) to deal with this kind of local modes. We call it Metropolized independence group sampler (MIPS). We first propose a new ϕ_g independent of old ϕ_g, say, from its prior distribution or an approximation of its conditional posterior distribution. Then, we sample all A_{ge} conditional on the new ϕ_g. The Metropolis-Hastings rule is used to decide whether to accept this move or not. To get a good proposal of A_{ge}, we use linear regression to get the least square estimate of A_{ge} and use it as the center of the proposal distribution of A_{ge}.
- We again use MIPS to deal with the strong correlation within the trend parameters $(a_{ge}, b_{ge}, c_{ge}, d_{ge})$ for a time series. The key is to propose

a new d_{ge} independent of the old d_{ge} and sample (a_{ge}, b_{ge}, c_{ge}) jointly conditional on the new d_{ge}, which is a multivariate normal distribution here.

- There are also strong correlations between λ_e and all A_{ge} of the same experiment e. We still use MIPS to perturb the MCMC chain. We propose a new λ_e independent of the old λ_e and sample all A_{ge} of the same experiment e conditional on the new λ_e. Similar to the MIPS moves for ϕ_g and A_{ge} of the same gene g, we used the least square estimate of A_{ge} to improve the proposal efficiency.

It should be noted that MIPS improves the mixing of the MCMC chain, especially at the initial state of the sampling, with an extra cost in computation. Our simulations indicated that this is a worthy effort. In meta-analysis, it is not unusual that different experiments support different values for a shared parameter. As a result, the shared parameter may have a multimodal distribution. In that case, strategies such as MIPS for making global moves are desirable.

ACKNOWLEDGEMENT

We thank Prof. Xiao-Li Meng for his helpful suggestions that led to a modification of the model in this paper. We are also grateful to Prof. Wing H. Wong and Dr. Xin Lu for their valuable comments. This research is supported in part by the NIH grant R01GM078990 and the NSF grant DMS-0706989. All R codes and fitting results are available upon request.

REFERENCES

Ahdesmaki, M., Lahdesmaki, H., Pearson, R., Huttunen, H., and Yli-Harja, O. (2005). Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics* **6**, 1, 117.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, 267–281. eds. B. N. Petrox and F. Caski, Budapest: Akademiai Kiado.

Bar-Joseph, Z., Siegfried, Z., Brandeis, M., Brors, B., Lu, Y., Eils, R., Dynlacht, B. D., and Simon, I. (2008). Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proc. Natl Acad. Sci. USA* **105**, 3, 956–961.

Chen, D., Toone, W. M., Mata, J., Lyne, R., Burns, G., Kivinen, K., Brazma, A., Jones, N., and Bahler, J. (2003). Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* **14**, 1, 214–229.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. Am. Stat. Assoc* **90**, 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis hastings output. *J. Am. Stat. Assoc* **96**, 270–281.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 1, 65–73.

de Lichtenberg, U., Jensen, L. J., Fausboll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* **21**, 7, 1164–1171.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868.

Futschik, M. E. and Herzel, H. (2008). Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis. *Bioinformatics* **24**, 8, 1063–1069.

Gelman, A., li Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 4, 711–732.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley, New York.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 1, 97–109.

Hertz-Fowler, C., Peacock, C. S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A. C., Rajandream, M.-A., and Barrell, B. (2004). GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucl. Acids Res.* **32**, suppl_1, D339–D343.

Ishida, S., Huang, E., Zuzan, H., Spang, R., Leone, G., West, M., and Nevins, J. R. (2001). Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol Cell Biol* **21**, 4684–4699.

Johansson, D., Lindgren, P., and Berglund, A. (2003). A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics* **19**, 467–473.

Klevecz, R. R., Bolen, J., Forrest, G., and Murray, D. B. (2004). From the cover: A genomewide oscillation in transcription gates DNA replication and cell cycle. *Proc. Natl Acad. Sci. USA* **101**, 5, 1200–1205.

Lange, K. L., Little, R. J. A., and Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Am. Stat. Assoc* **84**, 408, 881–896.

Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M., and Shapiro, L. (2000). Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290**, 2144–2148.

Liu, D., Umbach, D. M., Peddada, S. D., Li, L., Crockett, P. W., and Weinberg, C. R. (2004). A random-periods model for expression of cell-cycle genes. *Proc. Natl Acad. Sci. USA* **101**, 19, 7240–7245.

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing* **6**, 2, 113–119.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York.

Liu, J. S. and Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* **87**, 2, 353–369.

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes.

Biometrika **81**, 1, 27–40.

Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *J. Am. Stat. Assoc* **94**, 1264–1274.

Lu, X., Zhang, W., Qin, Z. S., Kwast, K. E., and Liu, J. S. (2004). Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucl. Acids Res.* **32**, 2, 447–455.

Lu, Y., Mahony, S., Benos, P., Rosenfeld, R., Simon, I., Breeden, L., and Bar-Joseph, Z. (2007). Combined analysis reveals a core set of cycling genes. *Genome Biology* **8**, 7, R146.

Luan, Y. and Li, H. (2004). Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics* **20**, 3, 332–339.

Marguerat, S., Jensen, T. S., de Lichtenberg, U., Wilhelm, B. T., Jensen, L. J., and Bahler, J. (2006). The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. *Yeast* **23**, 4, 261–277. 10.1002/yea.1351.

Menges, M., Hennig, L., Gruissem, W., and Murray, J. A. (2002). Cell cycleregulated gene expression in arabidopsis. *J Biol Chem* **277**, 41987–42002.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B., and Leatherwood, J. (2005). The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biology* **3**, 7, e225.

Peng, X., Karuturi, R. K. M., Miller, L. D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L.-S., Liu, E. T., Balasubramanian, M. K., and Liu, J. (2005). Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell* **16**, 3, 1026–1042.

Ptitsyn, A. A., Zvonic, S., and Gimble, J. M. (2007). Digital signal processing reveals circadian baseline oscillation in majority of mammalian genes. *PLoS Computational Biology* **3**, 6, e120.

Rustici, G., Mata, J., Kivinen, K., Lio, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., and Bahler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics* **36**, 809–817.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 2, 461–464.

Shedden, K. and Cooper, S. (2002). Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl. Acad. Sci. USA* **99**, 4379–4384.

Sherr, C. J. (1996). Cancer cell cycles. *Science* **274**, 1672–1677.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9**, 12, 3273–3297.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. Roy. Stat. Soc. Ser. B* **64**, 4, 583–616.

Tsiporkova, E. and Boeva, V. (2008). Fusing time series expression data through hybrid aggregation and hierarchical merge. *Bioinformatics* **24**, 16, i63–69.

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. *Science* **310**, 5751, 1152–1158.

Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977–2000.

Wichert, S., Fokianos, K., and Strimmer, K. (2004). Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* **20**, 5–20.

Willbrand, K., Radvanyi, F., Nadal, J.-P., Thiery, J.-P., and Fink, T. M. A. (2005). Identifying genes from up-down properties of microarray expression series. *Bioinformatics* **21**, 20, 3859–3864.

Zhao, L. P., Prentice, R., and Breeden, L. (2001). Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA* **98**, 5631–5636.

Zhou, C., Wakefield, J., and Breeden, L. (2005). Bayesian analysis of cell-cycle gene expression data. *UW Biostatistics Working Paper Series* Working Paper 276.

Department of Statistics
Harvard University
One Oxford Street
Cambridge, MA 02138
USA
E-mail: xfan@stat.harvard.edu
jliu@stat.harvard.edu

Department of Statistics
Chinese University of Hong Kong
Shatin, N.T.
Hong Kong
E-mail: xfan@sta.cuhk.edu.hk

Broad Institute of MIT and Harvard
7 Cambridge Center
Cambridge, MA 02142
USA
E-mail: spyne@broad.mit.edu

TABLE 1. *Summary of the ten experiments for the fission yeast cell cycle*

data set name	Rustici *et al.*					Peng *et al.*		Oliva *et al.*		
microarray type	spotted PCR array					spotted oligo array		spotted PCR array		
synchronization technique	elutriation			cdc25		elutriation	cdc25	elutriation		cdc25
experiment name	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
number of covered gene	4113	3921	4176	4281	4173	4263	4571	4543	4400	4727
number of time point (S_e)	20	20	20	19	18	33	38	33	50	52
time point frequency (min)	15	15	15	15	15	10	10	15-21	2-10	10-15

NOTE:

1. The data set Rustici *et al.* is downloaded from http://www.sanger.ac.uk/PostGenomics/S_pombe/projects/cellcycle/. Peng *et al.* is downloaded from http://giscompute.gis.a-star.edu.sg/∼gisljh/CDC/CDC_dnld_data.html. Oliva *et al.* is downloaded from http://publications.redgreengene.com/oliva_plos_2005/.
2. The downloaded data set Rustici *et al.* has been normalized on an array-by-array basis using an in-house normalization script, which performs three steps: masking bad spots, filtering lower quality spots, applying local window-based normalization. Peng *et al.* has filtered low intensity features (2-fold less than the background) and done LOWESS normalization within array. Oliva *et al.* has been normalized within array by the GenePix Pro software with default setting.
3. Elutriation experiments are done to wild-type fission yeast, where samples of uniformly sized cells are obtained. Because cell size is correlated with cell cycle stage, these cells are synchronized with respect to their position in the cycle. Cdc25 block-and-release experiments are done to the fission yeast strain carrying the temperature-sensitive cdc25-22 mutant gene, where cells are initially synchronized by blocking them at some particular cell cycle stage, then releasing them from the block and taking samples at different times.

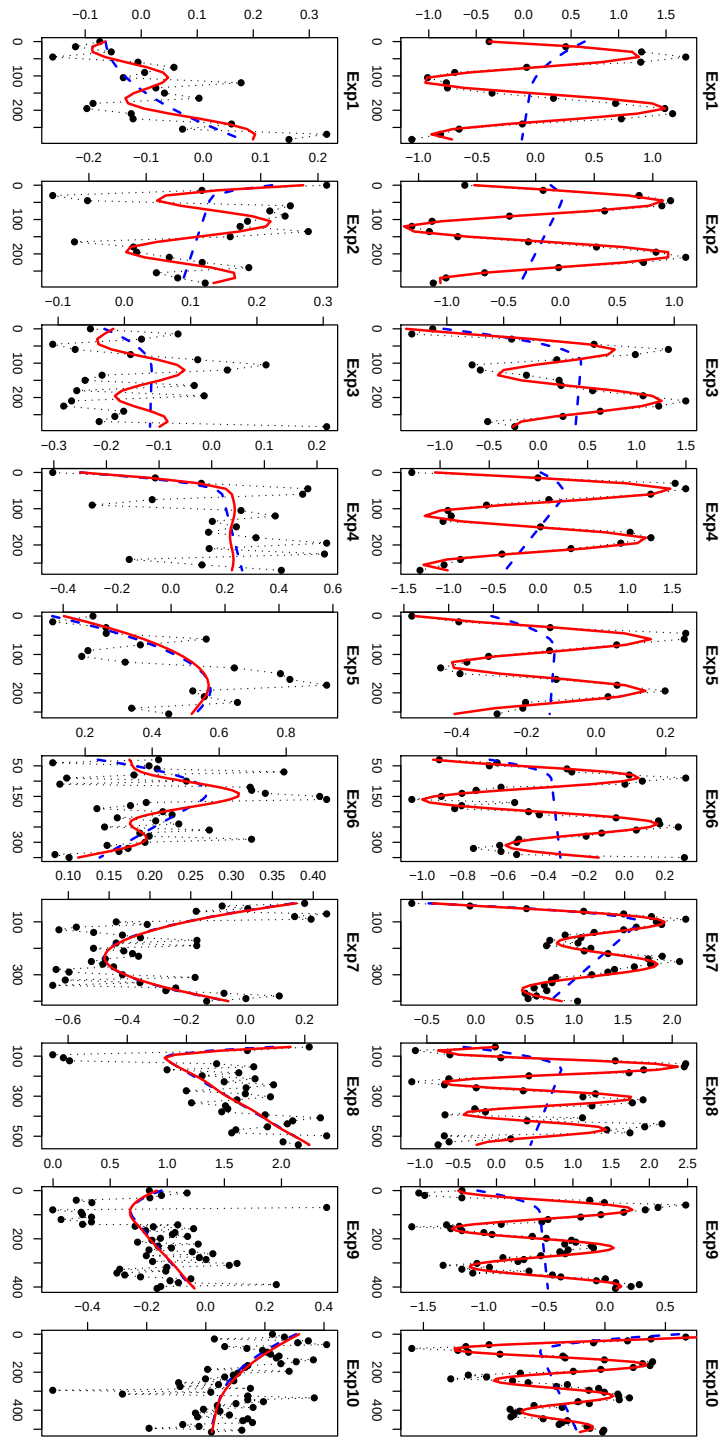

FIG 1. *Observed data and fitted mean curves for two samples of genes. For each sub-figure, the horizontal axis is the time (minutes) and the vertical axis is the gene expression value (log-ratio). The first row of sub-figures shows the ten time series for a known PE gene (SPAPYUG7.03C). The second row is for a stress response gene (SPAC23C4.09C), which is not regulated by the cell cycle. The bullet dots are the observed data. They are connected by dotted lines. The solid lines are the mean curves obtained by fitting the M_1 model to the data. The dashed lines are the mean curves obtained by fitting the M_0 model to the data. The details of model fitting are given in following text.*

TABLE 2
Correlation of different statistics and their classification results

Statistic	SNR	LPI	BIC^{01}
SNR	3599	3051	1967
LPI	-0.93	3086	1906
BIC^{01}	0.86	-0.83	2003

NOTE: The permuted data was used as background control. The lower-left part of the table shows the Spearman correlation between pairs of statistics. The numbers on the diagonal are the number of PE genes claimed by corresponding statistic. For SNR, we use a cutoff corresponding to FPR=0.002 for the two mean SNR density. For LPI, we also use the threshold corresponding to FPR=0.002. We use zero as the threshold for BIC^{01}. The upper-right part of the table show the number of PE genes claimed by a pair of statistics. Within them, 1898 genes are claimed by all three statistics.

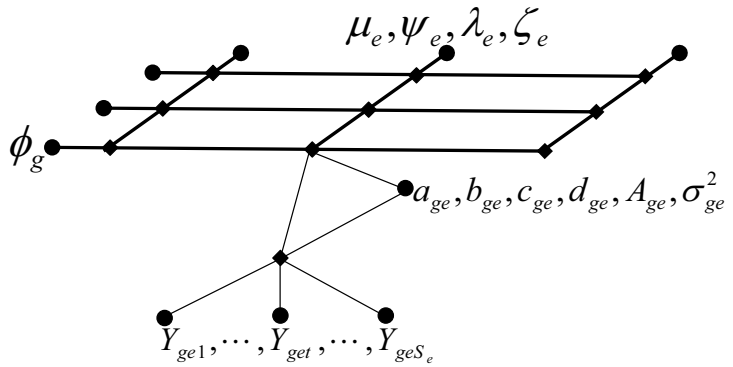

FIG 2. *Dependence structure of all variables. All links are undirected. Bullets represent a variable or a group of variables. Diamonds represent the dependence of the variables linked to it. Corresponding to the G-by-E matrix of time series, the main parameter structure can be visualized as a matrix, where each row corresponds to a gene-specific parameter ϕ_g and each column corresponds to experiment-specific parameters $(\mu_e, \psi_e, \lambda_e, \zeta_e)$. Each cell of the matrix corresponds to the variables specific to a time series. For example, all ϕ_g's are independent of each other conditional on all $(\mu_e, \psi_e, \lambda_e, \zeta_e)$; a time series is independent of all other time series conditional on the union of ϕ_g and $(\mu_e, \psi_e, \lambda_e, \zeta_e)$.*

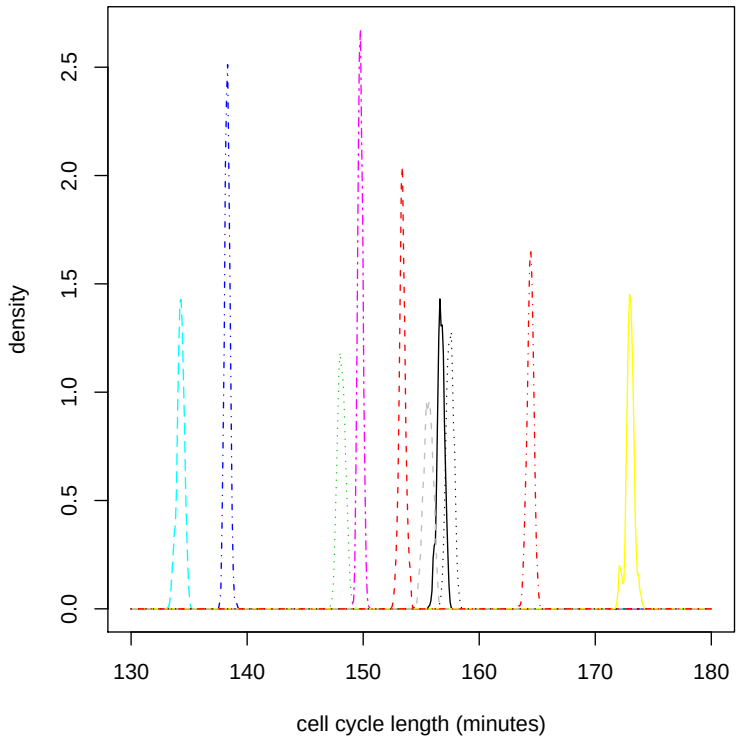

FIG 3. *Posterior distributions of all cell cycle lengths $2\pi/\mu_e$ (unit: minute). The posterior distribution of the cell cycle length is represented by the second half of the first 4000 iterations from 6 independent MCMC chains. The density curves of the ten $2\pi/\mu_e$ are drawn here.*

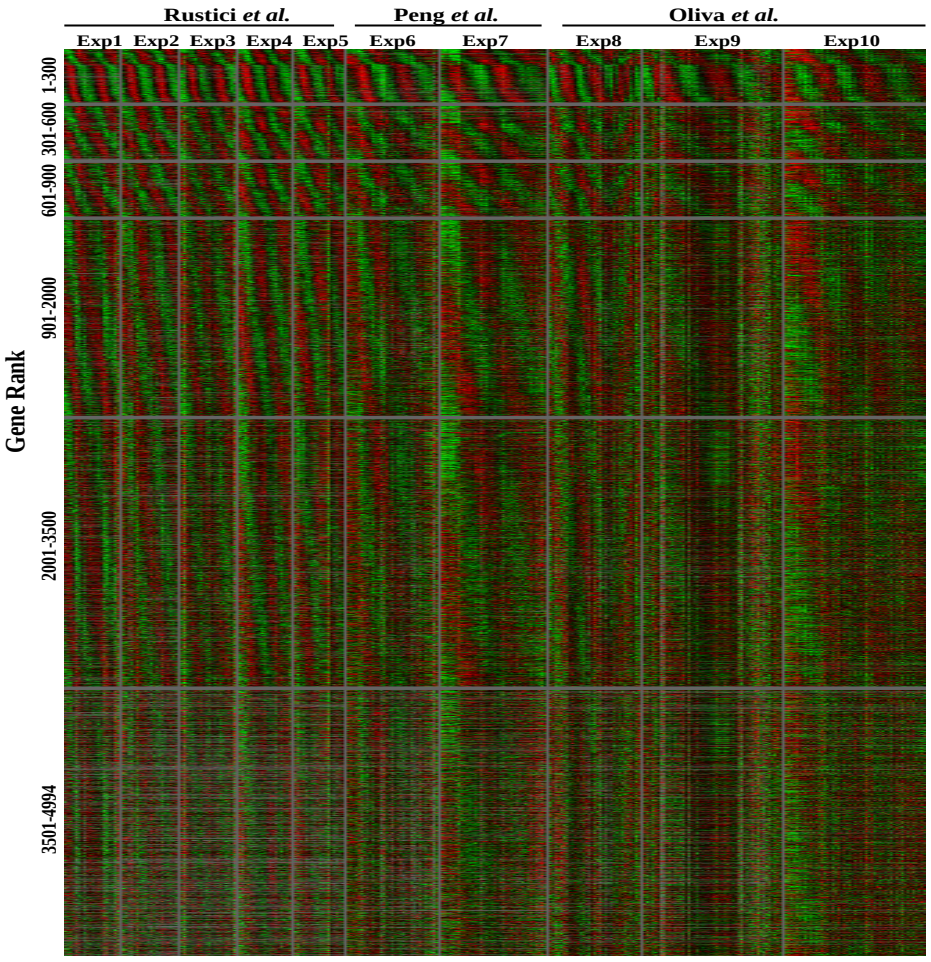

FIG 4. *Heatmap of all genes' time series data ranked by decreasing mean SNR value. Columns correspond to time points, which are grouped by experiment and sorted by time within each group. Rows correspond to genes, which are ranked by their mean SNR value and sorted by their mean peak times within each group. For example, the first row group contains the 300 genes with the highest mean SNR value from our combined analysis of all 10 experiments, and they are sorted by their relative phase* $\phi_g + \psi_1$ *within the group. Each time series is normalized to zero mean and unit variance for display. The heatmap is drawn by TreeView (Eisen* et al.*, 1998) with default setting. Red indicates up-regulation, green indicates down-regulation, black means no change of expression levels, and grey is missing data. It shows a periodic pattern for all gene groups.*

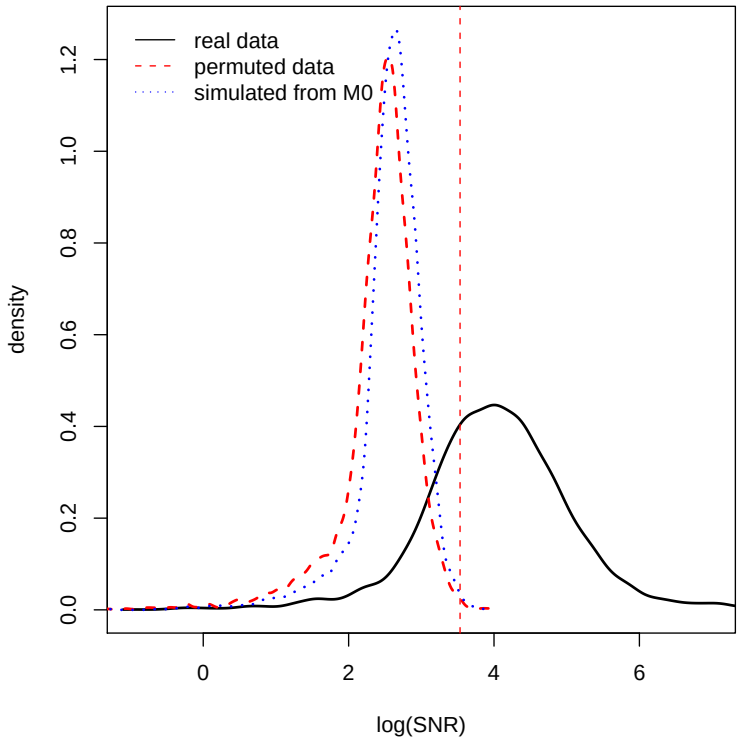

FIG 5. *Density comparison of SNR from the three data sets. The M_1 model is fitted to the real data, permuted data, and the data simulated from the M_0 model. For each gene, we get the posterior mean of the SNR statistic from the combined analysis. For each data set, we pool all genes together to get a kernel density estimate, which is shown in this graph. The vertical line indicates the threshold corresponding to FPR=0.002 in the permuted data, from which one can claim 3599 PE genes from the real data.*

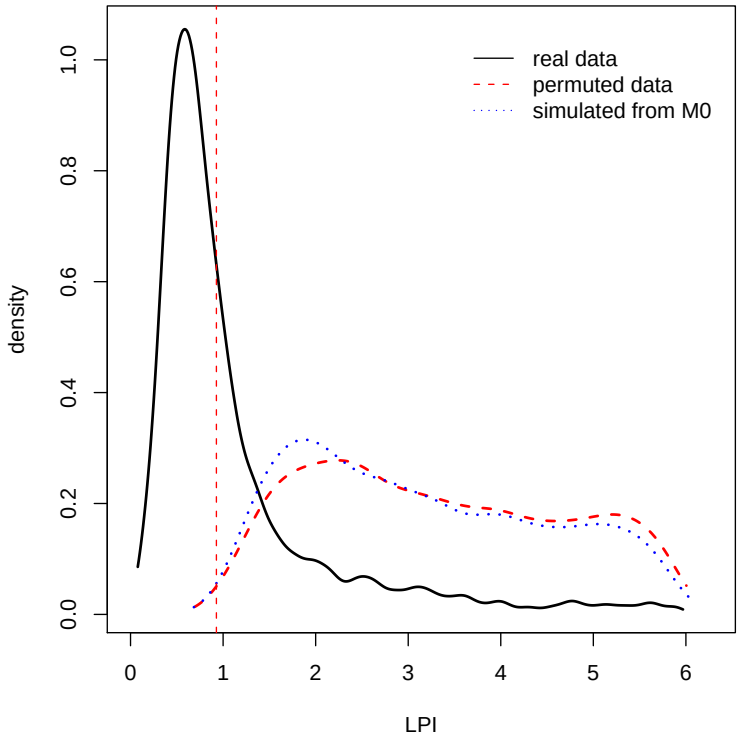

FIG 6. *Density comparison of LPI from the three data sets. LPI is defined as the Length of the 95% central Posterior Interval (LPI) for a gene's relative phase* $\phi_g + \psi_1$. *LPI measures our uncertainty about a gene's peaking time. The vertical line indicates the threshold corresponding to FPR=0.002 in the permuted data, from which one can claim 3086 PE genes from the real data.*

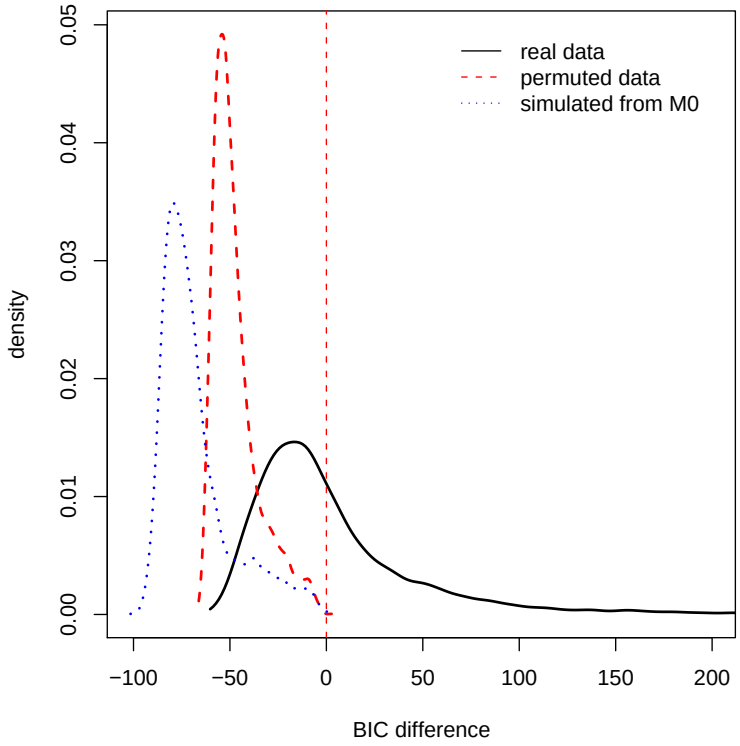

FIG 7. *Density comparison of BIC^{01} from the three data sets. Both the M_1 model and the M_0 model are fitted to the three data set. For each gene, BIC^{01} is calculated at the posterior mode to compare its fitting between the M_0 model and the M_1 model. The vertical line indicates the threshold $BIC^{01} = 0$, from which one can claim 2003 PE genes from the real data and only 1 gene from the permuted data.*

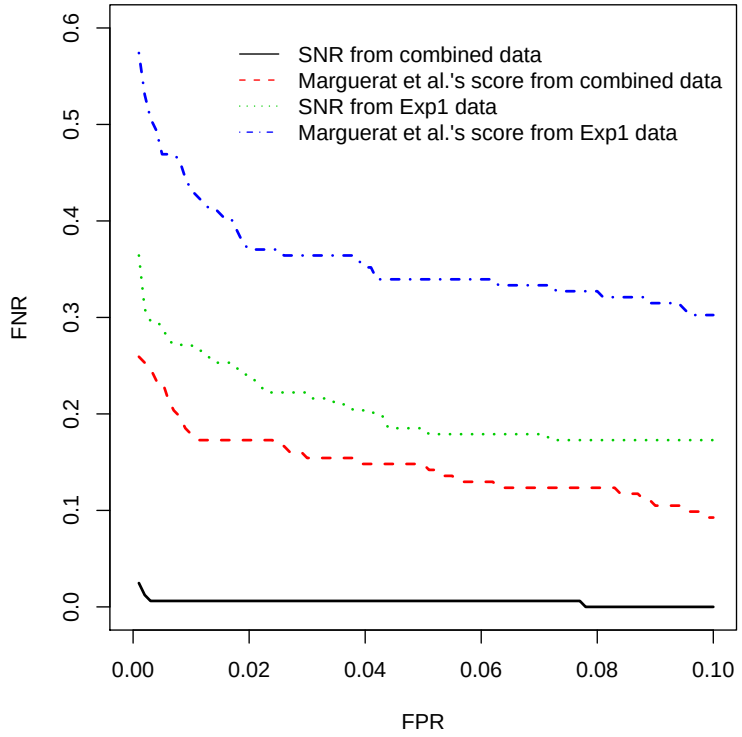

FIG 8. *Performance on the benchmark set. For each of the four methods listed in the figure legend, we plot FNR against FPR under various thresholds. For each threshold, the benchmark set of 162 PE genes is used to estimate FNR. The permuted version of the data is used to estimate FPR. A smaller under-curve area corresponds to a better classification performance for the benchmark set.*

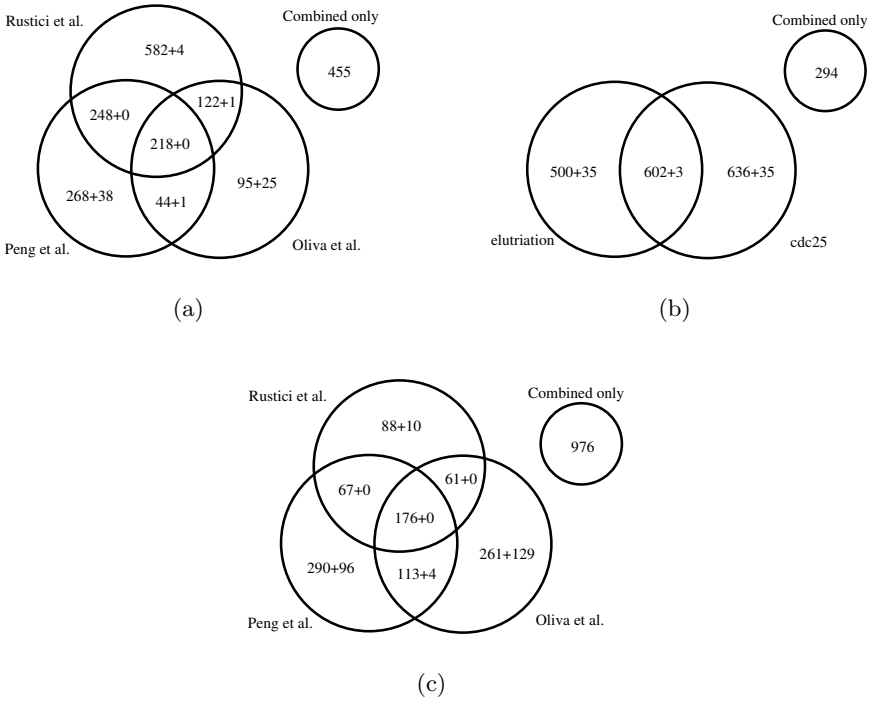

FIG 9. *Venn diagrams showing overlap between claimed PE genes from subsets of the data. Each gene set in all diagrams is compared with the result from the combined analysis that we did using our method. The number before the plus sign is the number of genes also claimed as periodically expressed by our combined analysis. The stand-alone circle represents the part which is reported only by the combined analysis. (A) Comparing the results from individual data sets using our method. (B) Comparing the results from two synchronization techniques using our method. (C) Comparing the results reported in original studies.*

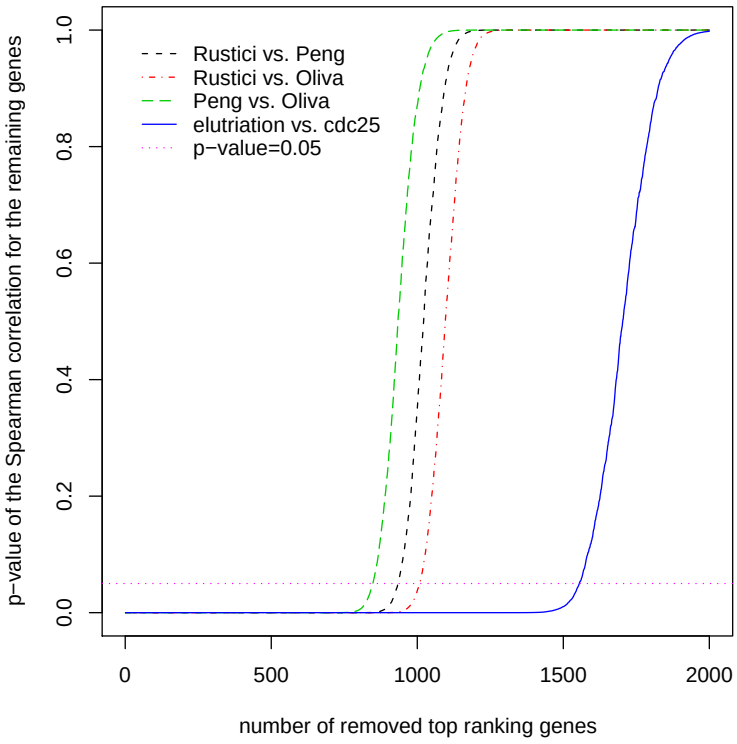

FIG 10. *Spearman correlation test for reproducibility between pairs of data sets. For each data set, we rank genes in the order of decreasing mean SNR value. For a pair of data sets, we define a gene's combined rank as its lower rank in the two data sets. We then sort genes according to the combined rank. Top genes in the resulted list are the overlapping part of the top ranking genes from individual lists, i.e., the reproduced part if we claim the same number of PE genes from individual lists. To test the reproducibility between a pair of data sets, we remove a certain number of the top ranking genes in the resulted list and then do Spearman correlation test for the remaining genes' mean SNR value. The p-value of this test will tell us whether the two data sets will still produce consistent gene ordering after we take out the most consistent genes. This graph shows the relationship between the number of removed top ranking genes and the Spearman correlation test p-value. Different curves correspond to different pairs of data set. The horizontal dotted line corresponds to the p-value threshold of 0.05. The solid curve and the horizontal dotted line crossed when the number of removed genes increases to 1554, which means the elutriation experiment data and the cdc25 experiment data will not produce significant overlapping after we deleted the 1554 top ranking genes.*